AD-A218 421

DTIC FILE COPY

2

MENTATION PAGE

Form Approved
OMB No. 0704-0188

i is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, ting and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this cing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson nd to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY (Leave bla...

2. REPORT DATE
August 24, 1989

3. REPORT TYPE AND DATES COVERED
FINAL REPORT, 01 Apr 87 to 31 Mar 89

4. TITLE AND SUBTITLE
PARAMETRIC MODELS FOR A_n: SPLITTING PROCESSES AND MIXTURES

5. FUNDING NUMBERS
AFOSR-87-0192

61102F 2304/A5

6. AUTHOR(S)
Bruce M. Hill

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Michigan
Department of Statistics
AnnArbor, MI 48109-1092

8. PERFORMING ORGANIZATION REPORT NUMBER
AFOSR-TR- 90-0212

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR/NM
Building 410
Bolling AFB, DC 20332-6448

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

DTIC
ELECTE
FEB 26 1990
S D
B

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A class of parametric models, called splitting processes, is defined, using de Finetti's concept of adherent mass. Such splitting processes give rise to complex mixtures of distributions. It is proved that the nonparametric Bayesian predictive procedure, A_n, of Hill (1968), holds exactly for a member of this class called a nested splitting process. It is also shown that the generalization of A_n, called H_n, to deal with ties, can hold exactly. A multivariate version of A_n, based upon the splitting processes, is proposed. Some general considerations concerning ties and adherent masses are discussed, as well as their connection with the Dirichlet process. These include the phenomenon by which in the Dirichlet process, the posterior predictive mass builds up at the observed points, while under A_n no mass is given to the observed points, and under H_n some but not necessarily all posterior predictive mass builds up at the observed points. A very general class of splitting processes is then defined, which allows for some of the adherent mass at a point to be replaced by an exact tie. It is proved that both the Dirichlet process of Ferguson and A_n can arise as different special cases of this general model.

14. SUBJECT TERMS

15. NUMBER OF PAGES
31

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	**18. SECURITY CLASSIFICATION OF THIS PAGE**	**19. SECURITY CLASSIFICATION OF ABSTRACT**	**20. LIMITATION OF ABSTRACT**
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	SAR

Parametric Models for A_n : Splitting Processes and Mixtures

Bruce M. Hill[*]

July, 1987
Revised August 24, 1989

Abstract

A class of parametric models, called splitting processes, is defined, using de Finetti's concept of adherent mass. Such splitting processes give rise to complex mixtures of distributions. It is proved that the nonparametric Bayesian predictive procedure, A_n, of Hill (1968), holds exactly for a member of this class called a nested splitting process. It is also shown that the generalization of A_n, called H_n, to deal with ties, can hold exactly. A multivariate version of A_n, based upon the splitting processes, is proposed. Some general considerations concerning ties and adherent masses are discussed, as well as their connection with the Dirichlet process. These include the phenomenon by which in the Dirichlet process, the posterior predictive mass builds up at the observed points, while under A_n no mass is given to the observed points, and under H_n some but not necessarily all posterior predictive mass builds up at the observed points. A very general class of splitting processes is then defined, which allows for some of the adherent mass at a point to be replaced by an exact tie. It is proved that both the Dirichlet process of Ferguson and A_n can arise as different special cases of this general model.

KEYWORDS: Bayesian nonparametric statistics; prediction.

1 Introduction

A_n and H_n were proposed by Hill (1968, 1988b) for Bayesian inference in the case of extremely vague a priori knowledge as to the

[*] This work was supported by the U. S. Air Force under grant AFOSR-87-0192, and by the National Science Foundation under grant DMS-8901234. The US government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon.

90 02 23 061

form of the underlying distribution, i.e., Bayesian nonparametric prediction and inference. Such weak knowledge might be described in terms of "data on a rubbery scale." For example, it is known that A_n holds exactly when the observations are only simply ordered, as discussed in the above references, and this suggest that it might hold approximately even when there is something more than an ordinal scale of measurement. Earlier, Fisher (1939, 1948) had suggested a version of A_n from the fiducial point of view, and Dempster (1963) had elaborated and made more precise this insight of Fisher. Berliner and Hill (1988) applied the A_n model to deal with censored data in connection with survival analysis. Hill (1980a) showed that A_n yields a robust form of Bayesian inference, and provides approximations to many real-world situations. Hill (1988b) gave a new subjective Bayesian argument for A_n, reviewed its history, and because of the minimal and realistic assumptions underlying it, proposed A_n as a basic solution to the problem of induction, as defined, for example, by Hume (1748). Also Lenk (1984) showed that A_n arises from use of a log-Gaussian distribution for an unknown probability density function, and discusses the relationship between A_n and use of the empirical distribution function.

In this article I shall attempt to provide further justification for A_n, showing that it arises from simple parametric models, called splitting processes, and can ordinarily be viewed as appropriate when the data arise from the process of sampling from complex mixtures of distributions. Although Hill (1968) proved that A_n cannot hold for countably additive distributions for any n, it is known from Jeffreys (1961, p.171) that A_1 and A_2 do hold for conventional parametric models, and from the work of Lane and Sudderth (1978, 1984) that A_n is coherent in the sense of de Finetti for all n. Because of its practical importance for Bayesian statistics, it is essential also to understand precisely how A_n, for all n, can arise from simple conventional statistical models.

In Section 2 we define two basic types of splitting processes, and prove that the nested splitting process satisfies A_n. Also a multivariate version of A_n is proposed. Section 3 discusses some subtleties involved in dealing with tied or grouped data, as in Berliner and Hill (1988), and proves that H_n can hold exactly. Then in Section 3 an even more general class of splitting processes is defined, and it is proved that both the Dirichlet process of Ferguson and A_n arise as special cases of this model. Section 4 makes a few concluding remarks. The primary focus of this article is on predictive inference, as in Aitchison and Dunsmore (1975), Geisser (1971, 1982, 1985).

2 Splitting Processes

In this section we shall propose an explicit parametric model for A_n. Let x_i, for $i = 1, \ldots, n$, be the data values obtained in sampling from a finite population, and let the $x_{(i)}$ be their ordered values in increasing order of magnitude. Let X_i be the corresponding pre-data random quantities, so that the data consist of the realized values, $X_i = x_i$, for $i = 1, \ldots, n$. In this article, by A_n we shall mean the following three assumptions:

1. The observable random quantities $X_1, \ldots, X_n$, are exchangeable. [1]
2. Ties have probability 0.
3. Given the data x_i, $i = 1, \ldots, n$, the probability that the next observation falls in the open interval $I_i = (x_{(i-1)}, x_{(i)})$, is $\frac{1}{n+1}$, for each i $= 1, \ldots, n+1$. By definition, $x_{(0)} = -\infty$, and $x_{(n+1)} = +\infty$, unless explicitly stated otherwise.

We begin by recalling that A_1 and A_2 can be obtained by the use of improper prior distributions on the location, and on the location and scale parameters, respectively, of a normal distribution. See Jeffreys (1961, p. 171), Hill (1968, p. 688). For example, in the case of A_1, if μ has an improper prior distribution represented by Lebesgue measure, and if the distribution of the error is N(0, 1), then given $X_1 = x_1$, the posterior distribution of μ is N(x_1, 1), and the posterior predictive distribution for X_2 is N(x_1, 2). Hence the posterior probability that $X_2 > x_1$ is 1/2. Similarly, in the case of unknown μ and σ, if these parameters are given the conventional improper joint prior distribution of Jeffreys, then A_2 holds. For $n > 2$, until now A_n had not been obtained, constructively, by

[1] In Hill (1968, 1988b) exchangeability was not included in the *definition* of A_n in order to include more general situations, such as partial exchangeability.

DTIC COPY INSPECTED

ssion For
GRA&I ☑
TAB □
nounced □
ification

By
Distribution/
Availability Codes
Dist | Avail and/or Special
A-1

means of parametric models and improper prior distributions. Lane and Sudderth (1978) proved an existence theorem to the effect that finitely additive distributions satisfying A_n for each n exist, but did not explicitly model such distributions. Here we give explicit parametric representations that can hold for all n.

The first step in our construction is to introduce the concept of adherent mass at a point. This is an extremely simple and useful concept, due to Bruno de Finetti (1974, p. 240), that arises in the finitely additive theory of probability. Before making precise definitions, we shall motivate this concept in connection with the joint distribution of two variables, X_1 and X_2, which will later represent the first stage in our iterative construction of a splitting process satisfying A_n. Let X_1 have distribution π, where π is any fixed distribution on the line. We now describe the conditional distribution of X_2, given $X_1 = x_1$. With probability 1/2, X_2 is given the distribution π; with probability 1/2, X_2 is *adherent* to x_1, with conditional probability 1/2 of being larger than x_1, conditional probability 1/2 of being smaller than x_1, conditional probability 0 of being equal to x_1, and with conditional probability 1 of being within any open interval containing x_1. Such adherence can be obtained as follows. Imagine that, given $X_1 = x_1$, $X_2 - x_1$ is equal to 1/K for some nonzero integer K, where K has a distribution symmetric about 0. If K has a diffuse finitely additive distribution on the integers, so that there is probability 1 that K is larger in absolute value than any finite constant, the result follows easily, since K must be some finite integer, so that $X_2 - x_1$ cannot be 0, and with probability 1, 1/K will be smaller in absolute value than any positive constant. The concept of adherence does not depend upon symmetry, althorgh this is the primary case of interest in this article. Also, one can place some positive mass exactly at the point, x_1, which will be discussed in Section 3 in connection with ties and the Dirichlet process of Ferguson.

Such distributions may at first sight appear rather exotic, but this is not really the case. They correspond to a situation where no possible measurement can differentiate between a value and 0, for example, even though the quantity in question is known not to be equal to 0. In this case neither empirically nor theoretically can one rule out such adherent distributions. Thus we may know that a particle has positive mass, but its mass may be so small that it is enormously beyond the powers of our technology to determine the exact value. It may only be possible in finite time to determine that the value is less than some specified positive ϵ. Indeed, looked at too finely, it may turn out that there is no fixed exact numerical value, but rather that the quantity in question is constantly fluctuating. Similarly, consider a large positive integer, for example the total number of subatomic entities in the universe, where it is assumed for the sake of argument that this quantity is well defined. Since any such entity may eventually turn out to be itself divisible, it is clear that in any specified finite time the most that can be done (apart from purely theoretical arguments) is to place a lower bound on such a quantity. From a subjective point of view, one might well have probability 1 that this integer, although finite, is larger than any number that has ever been specified. Such a K would be adherent at $+\infty$. Its reciprocal would be adherent at 0.

It is not necessary that one views such situations as holding exactly. Indeed, the primary purpose of the concept of adherence is merely to provide useful approximations and ways of thinking about very common situations. I think that clear understanding of the property of adherence is necessary in order to deal with ties and the grouping of data, such as in H_n, and in understanding the behaviour of the Dirichlet process. This will be discussed further in Remark 5 of the present section and in Section 3. For our purposes at present it suffices to observe that such finitely additive distributions are known to exist, so that the description we have given for generation of X_1 and X_2 is a coherent one in the sense of de Finetti, i. e., no Dutch book is possible. If desired, they can equally well be represented in terms of improper prior distributions. For example, a uniform weight of unity for each positive integer generates adherence at 0 for the reciprocal of such an integer. For the most part, we will use the language of the finitely additive theory, which is fully rigorous, and whose foundations were developed by de Finetti (1974) and L. J. Savage (1972). See also Dubins (1975), Schervish et al (1984), and Hill and Lane (1985). Rényi (1970) and Hartigan (1983) provide rigorous theories of improper prior distributions and conditional probability spaces.

Since most probabilists and statisticians accept the countably additive framework, and therefore might immediately reject such concepts as that of adherent mass, it may also be useful to point out that the axiom of countable additivity (or continuity) has never been justified other than by expediency. For example, in the book which founded the modern measure-theoretic treatment of probability, Kolmogorov (1950, p. 15) says:

> For infinite fields, on the other hand, the Axiom of Continuity, VI, proved to be independent of Axioms I-V. Since the new axiom is essential for infinite fields of probability only, it is almost impossible to elucidate its empirical meaning, as has been done, for example, in the case of Axioms I-V in 2 of the first chapter. For, in describing any observable random process we can obtain only finite fields of probability. Infinite fields of probability occur only as idealized models of real random processes. *We limit ourselves, arbitrarily, to only those models which satisfy Axiom VI.*[2] This limitation has been found expedient in researches of the most diverse sort.

Although expediency is important, it is hardly a matter of fundamental truth. For this reason I ask the indulgence of the reader to pursue further some of these ideas, even though at first glance they may seem unusual. The issues concerning countable additivity have some important implications for the theory and practice of statistics. For example, Ramakrishnan and Sudderth (1988) have shown that even in the simplest of all probability scenarios, that of flipping a fair coin, Borel's Strong Law does not hold in the finitely additive context. These authors show that with exactly the same joint distributions for all finite sequences, i. e., probability $1/2^k$ for any k-tuple of 0's and 1's, one can have the average converge everywhere to 0, converge everywhere to 1, or fail to converge everywhere. For the practice of statistics, these issues boil down to questions as to choice of approximations. We shall see in Section 3 that both the Dirichlet process and A_n can be seen as special cases of a very general class of splitting processes, with the Dirichlet process substituting exact ties for the adherent mass distributions.

[2] Author's italics.

We shall now make a few definitions which will enable us to operate with such adherent distributions, and to define a splitting process.

- Definition 1: A probability distribution is said to have adherent mass at a point (finite or infinite) if the infimum of probabilities of all open neighborhoods of the point is greater than the probability of the point itself. It is said to have a purely adherent mass at a point if it has an adherent mass at the point and the probability of the point itself is 0. Such language is also used for random quantities with such distributions.
- Definition 2: A random quantity is said to be negligible if it has a mass of unity adherent to 0.
- Definition 3: Two random quantities are said to be equivalent if their difference is negligible.
- Definition 4: A distribution is said to be diffuse at $+\infty$ if it has a purely adherent mass of unity at $+\infty$, diffuse at $-\infty$ if it has a purely adherent mass of unity at $-\infty$, and diffuse at ∞ if it has a purely adherent mass of 1/2 at each of $-\infty$ and $-\infty$, respectively. (When π is diffuse at ∞, and a random quantity X has distribution π, we shall sometimes say that X splits from ∞, or is generated from ∞. When X has a distribution for which all of the mass is adherent to a point. x_i, we shall sometimes say X splits from x_i.)

It follows immediately that a finite sum of negligible quantities is negligible, and that a diffuse distribution attaches probability 0 to any finite interval. Special diffuse distributions are used by some Bayesians to represent a form of ignorance. The improper uniform prior distribution for a location parameter, and for the logarithm of a scale parameter, as in Jeffreys (1961), are familiar special cases. These can be given a finitely additive interpretation as well. One can also strengthen the notion of diffuseness by requiring that the conditional distribution for a particular value, conditional on a finite set of values, be uniform. as in Hill (1980c).

We now prove a simple lemma that will add insight as to the nature of adherent mass, and be used in our proof of Theorem 1.

Lemma 1 *If X and Y are equivalent random quantities, then their distribution functions at a point z are identical, provided that neither random quantity has mass adherent at z.*

Proof:

Let $F(t) = Pr\{X \leq t\}$ and $G(t) = Pr\{Y \leq t\}$ be the distribution functions for X and Y, respectively. Let $Y = X + \epsilon$, where ϵ is negligible. Then partitioning the event $\{Y \leq z\}$ according to whether or not $X > z + \delta$, yields

$$Pr\{Y \leq z\} \leq inf_{\delta>0} Pr\{X \leq z + \delta\}.$$

Hence

$$G(z) - F(z) \leq inf_{\delta>0}[F(z + \delta) - F(z)].$$

If this infimum is positive, then the distribution of X has positive mass adherent at z. Reversing the roles of X and Y, we see that also

$$F(z) - G(z) \leq inf_{\delta>0}[G(z + \delta) - G(z)].$$

Thus if neither distribution has mass adherent at z, then both infimums must be 0, and then F(z) = G(z).

$\triangle$

As de Finetti (1974, p. 242) points out, it is preferable to define the distribution function for random quantities, not for either closed or open intervals (to obtain right or left continuity, respectively, in the countably additive theory), but rather to think of the distribution function as indeterminate at discontinuity points. This idea is consistent with the view that a mass exactly at a point may be, practically speaking, indistinguishable from a mass adherent at the point, in which case the value of the distribution function at the point should be viewed as indeterminate.

There are some subtleties that arise in the finitely additive theory that are worth mentioning explicitly. Although a mass purely adherent to 0 is for practical purposes indistinguishable from a mass exactly at 0, the two associated random quantities are not logically identical, since the first is certain *not* to be exactly 0. In dealing with such things we must therefore take greater care than is customary in the conventional countably additive theory. For example, strictly speaking, a random quantity with mass exactly at 0 would not be exchangeable with one having the same mass purely adherent at 0. One might nonetheless call such random quantities exchangeable up to negligible differences. See also Remark 8 below.

We now proceed to construct a splitting process. Let X_1 and X_2 be defined as before. Given $X_1 = x_1$ and $X_2 = x_2$, we generate X_3 as follows. With conditional probability 1/3, X_3 is generated according to π; with conditional probability 1/3, X_3 is generated from a symmetrical distribution purely adherent at x_1; and with conditional probability 1/3, X_3 is generated from a symmetrical distribution purely adherent at x_2. This procedure can be continued iteratively. After $X_i = x_i$, i $= 1, \ldots, n$, have been obtained, the conditional distribution of X_{n-1} is equally likely, with common probability 1/(n $+$ 1), to be generated from π or to have a symmetrical distribution purely adherent to each of the n distinct values, x_i, already generated. In other words, X_n is equally likely to split from each of the n $+$ 1 points, $\infty, x_1, \ldots, x_n$. The observations are generated sequentially in time, so that we can speak of X_i as the i^{th} point generated. Finally, joint distributions of the X_i are defined so as to be forward disintegrable (or strategic) in the sense of Dubins (1975), Lane and Sudderth (1984), i. e., probabilities for future observations can be evaluated as expectations of conditional probabilities, given previous observations. We call such a sequence $X_1, \ldots, X_n$, for any fixed π, a nested splitting process.

We shall assume, for simplicity, that the finitely additive distributions for π and the adherent mass distributions have been defined for all subsets of the line. By virtue of de Finetti's fundamental theorem of probability, it is always possible coherently to extend any partially defined coherent evaluation of probability to all subsets, de Finetti (1974, p. 111), Lad et al (1987). Finally, exchangeability in the finitely additive context will be defined in terms of equality of joint distributions, in the sense of equality of joint distribution functions evaluated at finite points, just as in the countably additive case. See, however, Remark 8 below.

Theorem 1 *For a nested splitting process, with π diffuse at ∞, A_n holds exactly. If π is any distribution with neither adherent nor positive mass at finite points, then exchangeability still holds, and ties have probability 0.*

Proof:

That ties have probability 0 follows immediately from the definition of pure adherence and the fact that π has no adherent or positive mass at finite points. That the conditional probabilities are in accord with A_n when π is diffuse may be seen as follows. Let $X_i = x_i$, for $i = 1, \ldots, n$, with all of these values distinct, and consider the conditional distribution of X_{n+1}. (Note that in the finitely additive theory all conditional distributions automatically satisfy the axioms of probability, as with full conditional probability distributions. See Dubins (1975) and Hill and Lane (1985).) Now let I_i be the open interval between $x_{(i-1)}$ and $x_{(i)}$, for $i = 1, \ldots, n-1$. First take i to be between 2 and n, so that the I_i are finite intervals. Since I_i is finite and π is diffuse, if X_{n+1} is generated from π, then there is probability 0 that X_{n+1} will fall in I_i. Similarly, unless X_{n+1} splits from either $x_{(i-1)}$ or from $x_{(i)}$, there is probability 0 that X_{n+1} will fall in I_i. Conditional upon X_{n+1} splitting from $x_{(i)}$, the probability that it falls in I_i is 1/2, and similarly if X_{n+1} splits from $x_{(i-1)}$. Since there are n + 1 equally likely possible sources for X_{n+1}, including π, it follows that the probability that X_{n+1} falls in I_i is exactly 1/(n + 1). When π is diffuse, this is also true if i = 1 or i = n + 1, in which case the interval I_i is semi-infinite. For example, if i = 1, then (ignoring events of probability 0) in order for X_{n+1} to be in I_1, it must be the case that either X_{n+1} splits from $x_{(1)}$, or else that it is generated from π. In the latter case, because π is diffuse at ∞, there is probability 1/2 that X_{n+1} will be smaller than $x_{(1)}$. This yields 1/(n + 1), as before, for the posterior predictive probability that X_{n+1} will be in I_1. Similarly for i = n + 1. This completes the proof that the conditional distribution for X_{n+1} is in accord with A_n when π is diffuse.

We now prove that $X_1, \ldots, X_{n+1}$, form an exchangeable sequence, for any π which has no adherent or positive mass at finite points.

By first conditioning on $X_1 = u$, and then using disintegrability to integrate with respect to u, we have, for $s_1 < s_2$,

$$\begin{aligned} Pr\{X_1 \le s_1, X_2 \le s_2\} &= \int_{-\infty}^{s_1} Pr\{X_2 \le s_2 \mid X_1 = u\}\pi(du) \\ &= \int_{-\infty}^{s_1} [1/2\ \pi(s_2) + 1/2]\pi(du) \end{aligned}$$

$$= 1/2[\pi(s_1)\,\pi(s_2)] + 1/2\,\pi(s_1),$$

where $\pi(s)$ is the mass attached to the closed interval from $-\infty$ to s by π. With a similar evaluation for the case $s_1 \geq s_2$, we obtain the joint distribution,

$$Pr\{X_1 \leq s_1, X_2 \leq s_2\} = 1/2[\pi(s_1)\,\pi(s_2)] + 1/2\,\pi(s_1 \wedge s_2),$$

where $s_1 \wedge s_2$ is the smaller of s_1 and s_2. This function is symmetric in its arguments, proving that X_1 and X_2 are exchangeable.

By conditioning on the first k variables and using disintegrability, similar evaluations can be made for the higher dimensional distributions. Let $F^{(k)}(s_1, \ldots, s_k)$ be the joint distribution function for the first k random quantities, for $k = 1, \ldots, n+1$. Then it is easily verified that

$$\begin{aligned} F^{(k+1)}(s_1, \ldots, s_{k+1}) &= 1/(k+1)\,[\sum_{i=1}^{k} F^{(k)}(s_1, \ldots, s_{i-1}, s_i \wedge s_{k+1}, s_{i+1}, \ldots, s_k) \\ &+ \pi(s_{k+1})\,F^{(k)}(s_1, \ldots, s_k)], \end{aligned}$$

where for i = 1 in the above sum we take $(s_1, \ldots, s_{i-1}, s_i \wedge s_{k+1}, s_{i+1}, \ldots, s_k) = (s_1 \wedge s_{k+1}, s_2, \ldots, s_k)$.

Using the iterative character of such functions, it is easy to see that the joint distributions are symmetric functions of their arguments, which proves exchangeability. In the diffuse case, the joint distribution functions are in fact constant at finite points. For k = 1 the constant is 1/2, for k = 2 it is 3/8. If c_k is the constant for a k-dimensional joint distribution, then $c_{k+1} = c_k[k + 1/2]/[k+1]$.

$\triangle$

A few remarks may be useful to understand the above construction.

- Remark 1: When π is not diffuse A_n does not hold exactly, since given that X_{n+1} is generated from π, the probabilities of the I_i depend upon π and the x_i. However, A_n still holds asymptotically as $n \rightarrow \infty$, in the following sense. Suppose that we take a union of k_n of the I_i, where $k_n \rightarrow \infty$. Since the probability that X_{n-1} is generated from π is only 1/(n + 1), the posterior predictive probability for such a union is asymptotically the same as under diffuse π.

- Remark 2: A diffuse π is adherent at ∞, attaching probability 1/2 to any semi-infinite interval, and probability 0 to any finite interval. In the case of known bounds for the data values, one or both of the infinite points of adherence can be replaced by finite points. For example, if it is known that all variables are positive, then we can put in points of adherence at 0 and at $+\infty$, with each being equally likely. In other words, given that the point is from π, it now has probability 1/2 of being within any neighborhood of 0, and probability 1/2 of being within any neighborhood of $+\infty$. Similarly, if there is a known upper bound for the observations, the point at $+\infty$ can be replaced by such an upper bound. In the case of survival analysis, as in Berliner and Hill (1988). the times from treatment to death are non-negative, so we use the lower bound of 0 in place of $-\infty$.

- Remark 3: In the finitely additive theory all that was done above would remain valid if we were to deal with distributions concentrated on the rationals, instead of the real numbers. Indeed, this would ordinarily be the more realistic case.

- Remark 4: In our definition of adherency, we could have allowed some positive mass to be placed exactly at the point. In this case some observations would be exactly tied, as in H_n. See Section 3. Also, it may be observed that the theorem remains true when symmetry of the adherent distribution of errors is weakened and replaced by the assumption only that it is equally likely that errors are positive or negative.

- Remark 5: A subtle but important point is that in the context in which we are working, all distances are relative. Suppose that X_2 has split from X_1, and X_3 has split from X_2. The realized values, the x_i, can be visualized as such that the distance between x_3 and x_2 is negligible compared to that between x_2 and x_1. In other words, the former distance can be microscopic relative to the latter distance, despite the fact that under the concept of adherence, one initially viewed it as certain that X_2

and X_1 would be extremely close. (Since x_2 cannot be exactly the same as x_1, it is only a matter of relative distances; there is no absolute meaning to the word 'close'.) For example, with respect to the distances, one can think of a planet circling a sun, and with some satellite circling the planet. The concept of adherence, as interpreted in a practical and approximate sense, can allow for some very natural and familiar kinds of relationships between points, and can deal simultaneously with both macroscopic and microscopic distances.

- Remark 6: A splitting process as defined above cannot be constructed exactly by human endeavours. For that matter, neither can a uniform distribution on a finite interval. However, one can obtain approximations to such uniform distributions and other continuous distributions. Such approximations can then be used, with care, to obtain appoximations to our splitting process. For example, in this spirit, let π be a Cauchy distribution. Define a primary point to be a point generated directly from π. For splits from a primary point such as x_1, let the error be normal with mean 0 and standard deviation 1. For splits from a secondary point, i. e., a point that has itself split from some primary point, let the error distribution be normal with mean 0 and standard deviation .01. For splits from a tertiary point, let the error distribution be normal with mean 0 and standard deviation .0001, etc.

- Remark 7: There is an interesting, but incorrect intuition about A_n that is worth discussing. The initial reaction that some have to A_n is that it is unreasonable because it gives the same weight to enormously long finite intervals and extremely short intervals. A simple answer to this objection is to point out that perhaps the reason that an interval is very long is because there is little mass in that region, and the reason that other intervals are short is because there is substantial mass nearby. The nested splitting model provides a framework for this second intuition. In it each point is, so to speak, the center of its own universe. The model implies that there will be a group of sparsely distributed primary points, and around each of these there will be a network of sparsely distributed secondary points, etc. Such secondary points appear close together when viewed from the perspective of their associated primary point, but they appear sparsely distributed when viewed from the perspective of the tertiary points. The process is self-similar in the sense that the microscopic network of points that have split from some common ancestor, or have split from descendants of that ancestor, has the same character, no matter at what level that ancestor occurs. There are connections here with some of the concepts of fractile geometry, Barnsley et al (1988), Mandelbrot (1982). For a nested splitting model, the intuitions that stem from a naive interpretation of Lebesgue measure are simply not appropriate.

- Remark 8: An example of David Lane (private communication) shows that exchangeability in the finitely additive case may have some surprising implications. For the nested splitting process, given x_1 and that X_2 splits from x_1, there is probability 1/2 that $\mid X_2 \mid > \mid x_1 \mid$; given x_1 and that X_2 splits from ∞, there is probability 1 that $\mid X_2 \mid > \mid x_1 \mid$; so given x_1, there is a probability of 3/4 that $\mid X_2 \mid > \mid x_1 \mid$. Integrating with respect to x_1, we obtain 3/4 for the unconditional probability that $\mid X_2 \mid > \mid X_1 \mid$, rather than 1/2, as one might have expected. This does not contradict exchangeability of the X_i in the sense we have defined it, which is the usual sense, but shows that in the merely finitely additive case exchangeability for the X_i does not imply exchangeability for the $\mid X_i \mid$. (In the countably additive case, exchangeability for the X_i *does* imply exchangeability for the $\mid X_i \mid$. To understand why this need not be true in the merely finitely additive case, observe that in this case the complete probability distribution is not determined by the probabilities of rectangle sets, and therefore not by the joint distribution function. The event $\mid X_2 \mid > \mid X_1 \mid$ is not a rectangle set.) Plainly there are a number of quite subtle issues that arise with regard to the precise definition of exchangeability in the finitely additive case. We chose to define exchangeability in terms of invariance of the joint distribution functions both because this is the familiar definition in the countably additive case, and also because the definition in the finitely additive case has not yet been given serious attention, so we did not want to get involved with such intricacies in the present article, which is about statistical inference. If one wishes, one can regard exchangeability in the sense of equality of joint distribution functions as a form of weak exchangeability. In this case we have only proven weak exchangeability of the X_i. However, for the practical purposes of statistical inference being discussed in this article, this would seem quite sufficient.

In Lane's example, note that the marginal distributions for the $\{X_i\}$ are all the same, but not the joint distributions. A similar phenomenon occurs in connection with (4) of Hill (1968, p. 679), since (4) implies that the $X_{(j_i)}$ all have the same marginal distribution, although it is certain that they are in strictly increasing order.

The theorem shows that the probabilities specified by A_n can be realized exactly in theory. In our construction of the splitting process the time order was relevant to the realization of the process, or creation of the data. For example, X_2 could have split from infinity or from the already realized x_1, which requires the existence of x_1 before the determination of X_2. But we have also proved that the process so engendered is exchangeable, which implies that ***probabilistically*** this time order is immaterial, since under exchangeability the joint distributions are invariant under permutations. For a related situation consider the discussion of the relationship between the Pólya urn model and the Bayes-Laplace model in de Finetti (1974, p. 220), or the discussion of 'contagion' in Feller (1971, p. 57). Although two processes may be structurally different, the expression of our probabilistic knowledge about them can be precisely the same.

Because of the fact that the sequence $x_1, \ldots, x_n$, generated by a splitting process is exchangeable, we can forget the time ordering for the purposes of statistical inference. Thus we can instead consider a population of values $X_1, \ldots, X_N$, that originated from a splitting process, but now is simply an existing population of numbers. By construction these values are necessarily distinct, so that the ordered values are $X_{(1)} < X_{(2)} < \cdots < X_{(N)}$. Of course, before the process is realized, one can visualize the process as creating a random distribution, in which the probability attached to a set is simply the proportion of X_i in the set. However, in the present inferential context we will imagine that the values have already been generated, but unobserved. In the subjective Bayesian theory, so long as there is no further information about the population values X_i, it is appropriate to use the same distribution after, just as before, they were generated. See Hill (1988a) for discussion. Now suppose a simple random sample of size n, without replacement, is taken from such a population, and the observed ordered values in the data are $x_{(1)} < x_{(2)} < \cdots x_{(n)}$, as in Hill (1968).

Because of the exchangeability, one can suppose that these values are actually the first n values created by the process, so that A_n, and indeed A_{N-1}, is automatically satisfied in sampling from a population $X_1, \ldots, X_N$ that is created by a nested splitting process. It was proved in Hill (1968, p. 688) that A_k implies A_j for $j < k$, so in fact A_j can hold for any $j < N$. If one generates an infinite number of points, then A_n holds for all n, as for example in Lane and Sudderth (1978). If we define θ_i to be the proportion of the unsampled population falling in the interval $(x_{(i-1)}, x_{(i)})$, and if the population size is infinite, then it follows from a result of Hill (1968, p. 686) that A_n for all n implies that $\underline{\theta}$ has the uniform Dirichlet distribution on the $(n+1)$ dimensional simplex, no matter what values the $x_{(i)}$ take on. There is a finitely additive version of de Finetti's theorem for exchangeable random quantities, which suggests that the usual interpretation in terms of an 'unknown' distribution, representing the limiting frequency of points in various sets, is still valid, although uniqueness of the de Finetti measure is lost. See de Finetti (1937), Hewitt and Savage (1955), Lane and Sudderth (1978), Savage (1972, p. 53), Diaconis and Freedman (1980, 1981), and Hill (1988b) for some related discussion.

There is a second basic type of splitting process closely related to the first that is worth mentioning. Let X_1 be generated as before, but instead of observing it, suppose that we observe Y_1, which differs by a negligible quantity from X_1. Given $Y_1 = y_1$, with probability 1/2 let Y_2 be purely adherent to X_1; and with probability 1/2, let Y_2 be generated by first generating X_2 from π, and then taking Y_2 to be purely adherent to X_2. Continue in this way. The data generated will consist only of the y_i values, with the x_i playing the role of unseen quantities, somewhat like conventional parameters. For this reason, notationally we will replace the X_i by μ_i for such a process, and think of the μ_i as conventional location parameters. For this process, after n points have been generated, with m distinct μ_i, the probability that the next point is from a new μ_i is taken as 1/(m + 1), instead of 1/(n − 1) as in the nested splitting process for the probability of a split from ∞.

The process generated in this way leads to an exchangeable sequence of observables, Y_i, in which ties have probability 0. The proof of exchangeability follows the same lines as in the theorem. However, this process is conceptually quite different from the nested splitting process, and does not satisfy A_n but rather a modified version of A_n. In the original process one generates a nested array. For example, if the second value splits from the first, and then the third from the second, we may visualize this as the third being a satellite of the second, and the second as a satellite of the first. In three dimensions, for example, one can think of a moon of a planet of a sun. (Of course, our points are on the line and are not in orbit, and so we might take the projections along some ray of the positions on a particular date of all bodies, relative to some fixed origin, as determining the variable of interest.) With the second process, and with Y_1, Y_2 and Y_3 all splitting from μ_1, we would instead visualize as the data the positions of three planets circling a sun, which would not be part of the data. We shall refer to the second process as a planetary splitting process. In this analogy each sun corresponds to a μ_i, and the y_j represent the positions of the planets. The nested splitting process can be viewed as generating a heirarchical random effects model, while the planetary splitting process generates a one-way random effects model, in the analysis of variance. See Lindley and Smith (1972) and Hill (1965, 1977, 1980b) for Bayesian inference in random effects models.

Both of these processes engender mixtures of populations. For the first process, we can classify as a 'type' all those points that originate by splitting from the same primary point. For the second process, we can classify as a 'type' all those points that originate from the same 'sun' or 'nucleus', i. e., with the same μ_i. The reason that a planetary splitting process need not satisfy A_n is because the interval between two y_i that come from the same μ_j may have smaller probability for including the next observation than an interval that corresponds to points from different μ_j. (The exact values for such probabilities depends upon further specification of the adherent distribution of errors, which we shall not go into here.) However, it is easy to see that a modified version of A_n is satisfied. Suppose that we have observed n points y_i. Under the assumptions of our model there would be extremely high probability that one can break these up into some number m of non-empty groups or clusters, each corresponding to the same sun or nucleus. Indeed, returning to the stellar example, no one would ordinarily have difficulty in deciding to which solar system a particular planet belongs. Suppose, for example, that there are m clusters, with n_i points belonging to the i^{th} cluster, where $n_i \geq 1$ and $\sum_{i=1}^{m} n_i =$ n. Instead of using the intervals I_i between the individual observations as originally defined, we take the intervals $\bar{I}_i$ between the ordered group averages, say the $\bar{y}_{(i)}$, $i = 1, \ldots, m$. Now it is easy to see that the next observation will satisfy A_m. Indeed the argument is precisely the same as that given in the proof of the theorem, but with n replaced by m.

We have seen that A_n can hold exactly. Since the adherent masses can be represented in terms of infinitely many different limiting distributions, the two splitting processes we have defined are not unique. It is an open question, however, as to whether there is a *basically* different model from the nested splitting process that generates A_n exactly.

Finally, it is interesting to note that the splitting processes we have defined can immediately be generalized to higher dimensional spaces, for example, to the surface of a sphere, 3-dimensional Euclidean space, higher dimensional versions of these spaces, and indeed to any surface or space whatsoever. One need only generate points from an appropriate distribution π, and then define adherency in an appropriate way, using, for example, some metric in the space under consideration. Such generalizations lead to multivariate versions of A_n. For example, in 2-dimensional Euclidean space, one can take π to be diffuse in the sense of attaching probability 1 to the complement of any finite open sphere, with a purely adherent mass of 1/4 at each of the four points at infinity; and the adherent distribution of mass at a point can be taken as spherically symmetric about the point, giving mass 1/4 to each of the 4 quadrants formed with the point as origin. In this case there would be probability 1 that the next observation will be within any open sphere about a point, given that it splits from that point. In n dimensions we would attach probability $1/2^n$ to each of the quadrants formed by a point as origin, given that a split occurs from that point, again use spherical symmetry, and take π to have a purely adherent mass of $1/2^n$ at each of the 2^n points at infinity. One can proceed similarly on the surface of a sphere, except that now the symmetry must be restricted to the surface of the sphere. For more general surfaces and spaces there may be other notions of diffuseness and symmetry that are of interest. Also, in Bayesian survival analysis, as in Berliner and Hill (1988), there are a variety of different ways to introduce a multivariate version of A_n to allow for covariates. This will be discussed further in a separate article.

3 Ties and the Dirichlet Process

Suppose now that a splitting process, either nested or planetary, generates $X_1, \ldots, X_N$, to form a random population consisting of N distinct values. Let $X_{(1)} < X_{(2)} < \cdots < X_{(N)}$ be the order statistics for this finite population. Let M be the random number of (non-empty) types or groups in the population, where two units are in the same group if they have a common primary ancestor for the nested process, and are in the same group if they have split from the same μ for the planetary process. In the general case two units belong to the same group if their values differ in their generation by negligible quantities in the sense of Definition 2. Let the i^{th} group have the random positive integer L_i of units. The units in this group will not have exactly the same value, but under the model their values are likely to be relatively close. Let $\bar{X}_i$ be the i^{th} group average, and let $\bar{X}_{(i)}$ be the i^{th} ordered group average, in increasing order of magnitude, for $i = 1, \ldots, M$. It is convenient to speak of the i^{th} group as having the common 'value' $\bar{X}_i$, although the actual values in the group are necessarily distinct. Note that for such splitting processes the random vector $\underline{L} = (L_1, \ldots, L_M)$ will necessarily be exchangeable, with $L_i \geq 1$, and $\sum_{i=1}^{M} L_i = N$.

Without using the notion of a splitting process, and with the types being defined by *exact* ties, rather than merely through adherency as in the present article, such a model was introduced in Hill (1968, Section 3) to generalize A_n for the case of ties. In that model, denoted by H_n, there is an arbitrary distribution for M, given N, $1 \leq M \leq N$, an arbitrary exchangeable distribution for $\underline{L}$, given M and N, and (4) of Hill (1968, p. 679) is satisfied. Sampling from such populations yields a posterior distribution for the remainder of the population, i. e., what is unseen, that generalizes the inference under A_n. Specific splitting processes, such as the nested or planetary models, are more general in that the ties need not be exact, but are less general in that they imply specific distributions for M, given N, and for $\underline{L}$, given M and N. In Hill (1968, 1980a) the Bose-Einstein distribution for $\underline{L}$, given M and N, was used for the purpose of inference about the percentiles of the population; in Hill (1968) it was used to obtain the posterior distribution of the number of distinct types in the population, with a uniform prior distribution for M, given N, and then generalized in Hill (1979) to a truncated negative binomial distribution for M, given N; and in Hill (1970, 1974) it was used to model Zipf's Law. Chen (1978, 1980) considered the general symmetrical Dirichlet-multinomial distribution for $\underline{L}$, given M and N. Lewins and Joanes (1984) used this same model. Although none of these articles was based upon the concept of a splitting process, it is easily shown that the nested splitting process yields the Bose-Einstein distribution, approximately, for the distribution of $\underline{L}$, given M and N, providing some justification for the original assumption.

We have proved that data generated according to the nested splitting process satisfies A_n exactly. Can we also so justify H_n? The answer is yes, since any process that generates A_n can automatically be used to generate data from H_n. For example, suppose the splitting model is used to generate data X_i, $i = 1, \dots, M$, that satisfies A_m. Now generate a random vector $\underline{S}$ of dimension M, that has any exchangeable distribution, with $S_i \geq 1$, $i = 1, \dots, M$, and $\sum_{i=1}^{M} S_i =$ N. Define a new population, with N units, to consist of S_i units having the value X_i, for $i = 1, \dots, M$. It is easy to verify that this new random population satisfies H_n. Thus both A_n and H_n are coherent models for the data. The question as to which is more appropriate raises some subtle and delicate questions concerning the meaning of ties and groups. (Note that it is possible to generalize the above construction, since it is not essential to take N and the S_i to be integer valued. In this case one can take the $S_i / \sum_{i=1}^{M} S_i$ to be arbitrary proportions.) We have therefore proved the following corollary to the Theorem:

Corollary 1 *If the property A_n holds for a process, then it is possible to modify the process so that H_n holds also, with an arbitrary exchangeable distribution for $\underline{L}$, given M and N.*

The original H_n implies that some observations will be exactly tied whenever $N > M$. In real world problems, ties can arise either from grouping or rounding of untied data, as discussed above in connection with splitting processes, or alternatively can arise from the nature of the data, as with integer valued data. In the survival analysis of Berliner and Hill (1988), the data are times to death after treatment. Time is usually taken to be a continuous variable, although some modern physicists dispute this, and argue that there is a basic unit of time, the chronon, of approximate magnitude 10^{-43} seconds, Whitrow (1980, p. 203). Clearly we are in no position to argue one way or another on this question. Indeed, at a very basic level, the nature of the measurement process itself is quite elusive and sophisticated. See Jeffreys (1957, Ch. 5-6), Luce and Narens (1987), Russell (1914), Whitehead (1920), and Whitrow (1980, Section 4.7). However, for practical purposes, the situation is much the same as that concerning the differentiation between a mass exactly at 0 and a mass partly or purely adherent to 0, since again it is beyond our technology to make measurements of sufficient precision. Of course, relative to other types of measurement, time can be measured extremely finely. If time were measured sufficiently accurately, it is unlikely that any two people would die at exactly the same time after treatment, even if time is truly discrete. Thus the untied model might be more realistic for such data. On the other hand, in practice time must be treated very crudely, and so our basic time unit may be weeks or months or even years. In this case it is largely immaterial whether we regard the underlying time variable as continuous, or discrete at a very refined level. Berliner and Hill based their analysis on A_n rather than H_n, and argued that when there are, for example, 3 deaths all grouped and called at 8 weeks, one can deal with this by imagining that these 3 deaths were actually at distinct times quite close to the nominal value 8. One can then use A_n to attach a probability of 2/(n + 1) to the interval between the smallest and the largest of these true (but unobserved) death times, and this yields a probability of about 2/(n + 1) for short intervals containing 8. This method is consistent with the results from use of the nested splitting model, as will now be explained.

Suppose that a finite population of N distinct values is generated by a nested splitting process, as described at the beginning of this section. Let the data consist of n distinct values x_i, with m groups or types, and $n_i \geq 1$ observations at $\bar{x}_{(i)}$, where $\bar{x}_{(i)}$ is the i^{th} ordered sample group mean, and $\sum_{i=1}^{m} n_i = n$. Note that under the splitting models, it is a priori probabilistically certain that one will be able to identify the various groups on the basis of their observed values, even without some other means of doing so. Because of the exchangeability, without loss of generality we can suppose that the x_i are the first n values generated from the splitting process. Then given the data, the conditional probability that the next observation is of the same type as those in the i^{th} ordered sample group is $n_i/(n + 1)$. This process is a generalized Pólya process, in which such a probability is a linear function of the observed number of units in a cell, and with in addition the possible creation of new types. See Zabell (1982) for relationships with the sufficiency postulate of W. E. Johnson (1932). It is also a generalization of the urn processes of Hill, Lane and Sudderth (1980, 1987). The Berliner-Hill method for dealing with ties gives very nearly the same result, namely, $(n_i - 1)/(n + 1)$. The slight difference arises because in the one case we are talking about whether the next observation is of the same type as the i^{th} ordered sample type, and in the other case we are discussing the mass to be attached to the interval between the smallest and largest of the n_i values that form the i^{th} ordered sample group.

Finally, it is interesting to compare the analysis from splitting models, or from H_n, with that from the Dirichlet process. This process can be derived, as in Blackwell and MacQueen (1973), as a generalized Pólya process, which is itself a special form of splitting process. In the notation of Blackwell and MacQueen, we have

$$Pr\{X_{n+1} \in \mathcal{B} \mid X_1, \ldots, X_n\} = \mu_n(\mathcal{B})/\mu_n(\mathcal{X}),$$

where $\mu_n = \mu + \sum_{i=1}^{n} \delta(X_i)$, $P(X_i \in \mathcal{B}) = \mu(\mathcal{B})/\mu(\mathcal{X})$, $\delta(x)$ denotes the unit measure concentrating at x, and $\mathcal{X}$ is the space of observations.

Now generalize my original nested splitting process so as to include an additional parameter η_n, for the probability that the next observation is from π, and with equal probability $(1-\eta_n)/n$ that the next observations splits from each of the n realized x_i. Then for any open interval $\mathcal{B}$, my model yields

$$Pr\{X_{n+1} \in \mathcal{B} \mid X_1, \ldots, X_n\} = \pi(\mathcal{B}) \times \eta_n + [C_n(\mathcal{B}) + 1/2\, D_n(\mathcal{B})] \times (1-\eta_n)/n,$$

where $C_n(\mathcal{B})$ is the number of observations amongst the first n that lie in $\mathcal{B}$, and $D_n(\mathcal{B})$ is the number of x_i that are on the boundary of $\mathcal{B}$. With $\eta_n = \mu(\mathcal{X})/[n + \mu(\mathcal{X})]$, and for $D_n(\mathcal{B}) = 0$ this is identical with the probability as given by equation (2) in the Blackwell-MacQueen representation of the generalized Pólya process. For $\mu(\mathcal{X}) = 1$, we have my original splitting process. Note that if $\mu(\mathcal{X}) = \infty$, then the above predictive probability is simply $\pi(\mathcal{B})$.

If we now make one further generalization then both the nested splitting process and the Dirichlet process become special cases of a single very general process. Define $\tau_{i,n}$ to be the probability that the next observation ties $x_{(i)}$, given the first n observations. Given that X_{n+1} splits from $x_{(i)}$ but does not tie $x_{(i)}$, let the mass $1 - \tau_{i,n}$ be symmetrically adherent to $x_{(i)}$. In my original construction $\tau_{i,n} = 0$ for each n and $i = 1, \ldots, n$, and $\eta_n = 1/(n+1)$. To obtain the Dirichlet process of Ferguson (1973, p. 209) with parameters α and $M = \alpha(\mathcal{X})$, we need only set $\tau_{i,n} = 1$ for each n and $i = 1, \ldots, n$, take $\pi = \alpha/M$, and $\eta_n = \alpha(\mathcal{X})/[n + \alpha(\mathcal{X})]$. In this case $D_n(\mathcal{B}) = 0$ in the above equation, which holds for all $\mathcal{B}$, and is identical with equation (2) of Blackwell and MacQueen. Note that if we thus choose the parameters so as to yield the Dirichlet process, and if further we assume countable additivity for the sequence of variables that are generated by the process, then the process is identical with that of Blackwell and MacQueen. Thus both the Dirichlet process and

A_n can be seen as quite different cases of such generalized splitting processes. We state these results as a theorem.

Theorem 2 *Let $X_i, i = 1, \ldots, n, \ldots,$ be a generalized splitting process with parameters η_n and $\tau_{i,n}$. Then for $\eta_n = 1/(n+1)$, and $\tau_{i,n} = 0$ for $i = 1, \ldots, n$, the process is a nested splitting process. For $\eta_n = \alpha(\mathcal{X})/[n - \alpha(\mathcal{X})]$ and $\tau_{i,n} = 1$ for $i = 1, \ldots, n$, and under the assumption of countable additivity, the process is a Ferguson Dirichlet process with parameter α.*

Of course if the process is to be countably additive, we must take $\tau_{i,n} = 1$, since adherent mass distributions cannot occur in that framework. Every countably additive model is necessarily finitely additive, but the requirement of countable additivity forces one to rule out certain parameter values in the construction of the generalized splitting processes. It should be observed that this requirement also rules out conventional improper prior distributions for parameters, since such distributions cannot be represented as *proper* countably additive distributions. Yet such prior distributions provide standard and useful approximations in ordinary parametric Bayesian statistics. I believe that the same is true here. In fact, it is well known that classical non-Bayesian inferential devices, such as confidence procedures for Gaussian distributions, correspond in the Bayesian framework to precisely such improper prior distributions. (It follows from the continuity theorem of de Finetti (1974, p. 132) that coherency is always preserved under passages to the limit. The finitely additive distributions that we employ in this article, such as the diffuse distribution π and the adherent mass distributions at the points, can all be obtained as limits of proper distributions, and these limits can all be equally well represented as improper distributions.)

The primary problem with the standard Dirichlet process is that with high probability it yields data for which the posterior predictive mass piles up at what was observed. This seems unrealistic, especially from a predictive point of view. In fact, in the words of Ferguson (1973, p. 210): "There are disadvantages to the fact that P chosen by a Dirichlet process is discrete with probability one. These appear mainly because in sampling from a P chosen by a Dirichlet process, we expect eventually to see one observation exactly equal to another." This is precisely what A_n avoids, since all the posterior predictive probability is placed on the open intervals between successive order statistics; while H_n is a more flexible procedure, which allows for various degrees of tied or nearly tied data. In the H_n model, it is the posterior distribution of M, given N, that determines the extent to which future data will be tied, as can be seen by integrating equation (11) of Hill (1968, p. 683) with respect to this posterior distribution. For example, if M is believed to be sufficiently large, given the data, then the posterior probability for a tie becomes small; if M = N, then ties cannot occur.

Of course, one might object that under the adherence assumption, taken literally, one expects that the observations will be extremely close together, and this is qualitatively similar to the situation for the Dirichlet process. In a certain sense this is true, but as discussed earlier in Remark 5, the word 'close' has no absolute meaning. If we look at different planets clustered around different suns, we have data for which the distances between objects in the same solar system are negligible compared to distances between different solar systems. Yet we would not ordinarily call our own planet close to our sun. This highlights the essential relativity of all such considerations. In any case predictions based upon A_n and H_n are quite different from those based upon the Dirichlet process. Although an interesting and important idea, of which the present theory can be regarded as a generalization, the standard Dirichlet process does not seem to allow for the flexibility of splitting processes, including the various senses in which A_n and H_n can be approximated by different types of splitting processes. For example, π and the distributions for the errors can be taken to be proper distributions, such that the error distributions are tightly concentrated relative to π. Although in our proof of Theorem 1 we employ diffuse distributions and adherent masses, we view these as only idealizations that provide us with insight as to more realistic situations that involve approximations. In the same way it is useful to have a concept of a circle and a sphere, but without pretending that various bodies (such as the planets) are exactly spherical. In the eloquent words of B. Mandelbrot (1982), "clouds are not spheres, mountains are not cones, coastlines are not circles, and bark is not smooth, nor does lightning travel in a straight line."

4 Concluding Remarks

The theorem of Section 2 shows that A_n, and therefore as an immediate corollary, also H_n can hold exactly. According to the usual methodology of statistics, both Bayesian and non-Bayesian, to justify their use in practice one would have to argue on an *a priori* basis, that the models that give rise to these procedures are appropriate in the context of the specific example wherein their use is contemplated. This is meant in the same sense in which one attempts to justify the use of the Gaussian distribution on the basis of various considerations, such as the central limit theorem. However, it is well known that in practice no such arguments are ever more than suggestive as to the possible appropriateness of a normality assumption. For example, Poincaré (1912, p. 171) states in connection with this distribution, "Tout le monde y croit cependant, me disait un jour M. Lippmann, car les expérimentateurs s'imaginent que c'est un théoréme de mathématiques, et les mathématiciens que c'est un fait expérimental," or "everybody believes in the law of errors, the experimenters because they think it is a mathematical theorem, and the mathematicians because they think it is an experimental fact." See also Hill (1969, 1988b).

It is clear that even apart from questions concerning adherent mass, and diffuseness of π, which are of course only meant as approximations, the nested and planetary splitting models are also at best only suggestive as to the possible appropriateness of A_n and H_n in practice. It is not conceivable that one could ever *prove* that such models (or any other, such as the Gaussian) are exactly true. What is needed for the purpose of the practitioner is instead a heuristic form of reasoning which allows him to use his considered judgment as to why one or another model might be roughly applicable in various kinds of examples. In my opinion A_n and H_n find their best justification for the practitioner in connection with sampling from complex mixtures of distributions, and Bayesian data analysis. Splitting processes generate such complex mixtures of distributions, for example, with each primary point of the nested process, or each μ of the planetary splitting process, serving as a component of a mixture. Sampling from a real-world finite population that is a complicated mixture of many component distributions gives rise to data for which I believe that A_n and H_n can provide useful approximations. The approach to H_n via mixtures also turns out to be intimately connected to the Bayesian analysis of random effects models.

In conclusion, we have here constructed splitting models that yield A_n and H_n exactly, have discusssed in the cited articles how they arise as approximations, and have discussed the relationship with the Dirichlet process of Ferguson. A_n and H_n appear often to be appropriate, apart from situations where there is explicit and substantial knowledge as to the form of the underlying distribution. They are in fact coherent versions of the conventional use of the empirical distribution function. When one uses the latter for predictive purposes, one pretends that the next observation is certain to tie one of the previous values. This is plainly unreasonable, and A_n and H_n allow one to drop such a pretence, while preserving the advantages of using a diffuse prior distribution for the values X_i, if one wishes. Furthermore, the more complexity and real-world character that a problem has, the more these methods seem to be favored over other methods of inference. It is my personal opinion that even in cases where there is some strong parametric knowledge, use of A_n or H_n would ordinarily be preferable, unless sample sizes are quite small. Thus, when sample sizes are sufficiently large, one can be virtually certain that any conventional parametric model would be inadequate. And even if the data were from such a model, the results from an analysis based upon that model would largely agree with those based upon A_n, anyhow, since both would tend to agree with the empirical distribution function for intervals containing several observations. The main situation where one might want to depart substantially from A_n and H_n is perhaps with very small sample sizes, and very detailed and precise a priori knowledge as to the underlying distribution. Even here, it would not ordinarily be because one particularly believes in the truth of the parametric model, but rather because use of the model allows convenient smoothing. When n is small, one may wish to do more smoothing than A_n allows, in order to get more precise results from the posterior distribution. This becomes a question of the utility of the model. See Dickey and Kadane (1980).

A_n was originally suggested from a fiducial point of view. It also has a confidence/tolerance interpretation. It is simple, intuitive, coherent, and has several subjective Bayesian interpretations and justifications. I hope that it will be used more widely by practitioners than has hitherto been the case.

REFERENCES

Aitchison, J., and Dunsmore, I. R. (1975), *Statistical Prediction Analysis*, Cambridge University Press.

Barnsley, M. F., Devaney, R. L., Mandelbrot, B. B., Peitgen, H.-O., Saupe, D., Voss, R. F. (1988). *The Science of Fractile Images*, Springer-Verlag.

Berliner, L. Mark., and Hill, Bruce M. (1988), "Bayesian nonparametric survival analysis," *Journal of the American Statistical Association, 83*, 772-784 (with discussion).

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson distributions via Pólya urn schemes," *The Annals of Statistics*, 1, 353-355.

Chen, Wen-Chen (1978), *On Zipf's Law*, University of Michigan Doctoral Dissertation.

Chen, Wen-Chen (1980), "On the weak form of Zipf's law," *Journal of Applied Probability*, 17, 611-622.

De Finetti, B. (1937), "La prévision: ses lois logiques, ses sources subjectives," *Annales de l'Institut Henri Poincaré*, 7, 1-68.

De Finetti, B. (1974). *Theory of Probability, Vol. 1*, London: John Wiley & Sons, Inc.

Dempster, A. P. (1963), "On Direct Probabilities," *Journal of the Royal Statistical Society B, 25*, 100-114.

Diaconis, P., and Freedman, D. (1980), "Finite exchangeable sequences," *The Annals of Probability, 8*, 745-764.

Diaconis, P., and Freedman, D. (1981), "Partial exchangeability and sufficiency," *Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions*, 205-236.

Dickey, J., and Kadane, J. (1980), "Bayesian decision theory and the simplification of models," in *Evaluation of Econometric Models*, J. Kmenta and J. Ramsey, eds., Academic Press, 245-268.

Dubins, L. (1975), "Finitely additive conditional probabilities, conglomerability, and disintegrations," *Annals of Probability, 3*, 89-99.

Feller, W. (1971), *An Introduction to Probability Theory and its Applications, Volume 2, Second Edition*, New York: John Wiley Sons.

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics,* 1, 209-230.

Fisher, R. A. (1939), "Student,"*Annals of Eugenics, 9,* 1-9.

Fisher, R. A. (1948), "Conclusions Fiduciare,"*Annales de l'Institut Henri Poincaré,* 10, 191-213.

Geisser, S. (1971), "The Inferential Use of Predictive Distributions." In *Foundations of Statistical Inference,* V. P. Godambe and D. A. Sprott, eds., Toronto: Holt, Rinehart and Winston, 456-469.

Geisser, S. (1982), "Aspects of the Predictive and Estimative Approaches in the Determination of Probabilities," *Biometrics,* 38, Supplement, 75-93, (with discussion).

Geisser, S. (1985), "On the Prediction of Observables: A Selective Updata,"(with discussion), in *Bayesian Statistics 2,* J. M. Bernardo, M. H. Degroot, D. V. Lindley, A. F. M. Smith eds., North-Holland, Valencia University Press.

Hartigan, J. (1983), *Bayes Theory,*, New York: Springer-Verlag.

Heath, D., and Sudderth, W. (1976), "de Finetti's theorem for exchangeable random variables," *The American Statistician, 30,* 188-189.

Hewitt, E., and Savage, L. J. (1955), "Symmetric measures on cartesian products," in *The Writings of Leonard Jimmie Savage-A Memorial Selection,* Published by The American Statistical Association and The Institute of Mathematical Statistics, 1981, 244-275.

Hill, B. M. (1965), "Inference about variance components in the one-way model," *Journal of the American Statistical Association,* 58, 918-932.

Hill, B. M. (1968), "Posterior distribution of percentiles: Bayes theorem for sampling from a finite population," *Journal of the American Statistical Association,* 63, 677-691.

Hill, B. M. (1969), "Foundations for the theory of least squares," *Journal of the Royal Statistical Society, Series B, 31,* 89-97.

Hill, B. M. (1970), "Zipf's law and prior distributions for the composition of a population," *Journal of the American Statistical Association,* 65, 1220-1232.

Hill, B. M. (1974), "The rank frequency form of Zipf's law," *Journal of the American Statistical Association,* 69, 1017-1026.

Hill, B. M. (1977), "Exact and approximate Bayesian solutions for inference about variance components and multivariate inadmissibility," in *New Developments in the Application of Bayesian Methods,* A. Aykac and C. Brumat, eds., North Holland, Chapter 9, 129-152.

Hill, B. M. (1979), "Posterior moments of the number of species in a finite population, and the posterior probability of finding a new species," *Journal of the American Statistical Association, 74,* 668-673.

Hill, B. M. (1980a), "Invariance and robustness of the posterior distribution of characteristics of a finite population, with reference to contingency tables and the sampling of species." In *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of Harold Jeffreys,* A. Zellner, ed., North-Holland, 383-395.

Hill, B. M. (1980b). "Robust analysis of the random model and weighted least squares regression," in *Evaluation of Econometric Models*, J. Kmenta and J. Ramsey, eds., Academic Press, 197-217.

Hill, B. M. (1980c). "On finite additivity, non-conglomerability, and statistical paradoxes," (with discussion) in *Bayesian Statistics*, J. M. Bernardo, M. H. Degroot, D. V. Lindley, A. F. M. Smith, eds., University Press: Valencia, Spain, 39-66.

Hill, B. M., (1988a). "A theory of Bayesian data analysis," to appear in *Bayesian Analysis in Econometrics and Statistics: Essays in Honor of George Barnard*, S. Geisser, J. Hodges, S. J. Press, A. Zellner, eds., North-Holland, 383-395.

Hill, B. M. (1988b), "De Finetti's theorem, induction, and A_n, or Bayesian nonparametric predictive inference," in *Bayesian Statistics 3*, J. M. Bernardo, M. H. Degroot, D. V. Lindley, A. F. M. Smith, eds., Oxford University Press, 211-241 (with discussion).

Hill, B. M. and Lane, D. (1985). "Conglomerability and countable additivity," *Sankhyá, 47, Series A*, 366-379.

Hill, B. M., Lane, David, and Sudderth, William (1980), "A strong law for some generalized urn processes," *The Annals of Probability*, 8, 214-226.

Hill, B. M., Lane, D., and Sudderth, W. (1987), "Exchangeable urn processes," *The Annals of Probability*, 15, 1586-1592.

Hume, David (1748). *An Enquiry Concerning Human Understanding*, London.

Jeffreys, H. (1957), *Scientific Inference*, Second Edition, Cambridge University Press.

Jeffreys, H. (1961), *Theory of Probability*, Third Edition, Oxford at the Clarendon Press.

Johnson, W. E. (1932), "Probability: the deductive and inductive problems," *Mind, 49*, 409-423. [Appendix on pages 421-423 edited by R. B. Braithwaite].

Kingman, J. F. C. (1975), "Random discrete distributions," *Journal of the Royal Statistical Society, Series B, 37*, 1-22 (with discussion).

Kolmogorov, A. N. (1950), *Foundations of Probability*, New York: Chelsea Publishing Co.

Lad, F., Dickey, J., and Rahman, M. (1987), "The fundamental theorem of prevision," Technical Report No. 506, University of Minnesota, School of Statistics.

Lane, D., and Sudderth, W. (1978), "Diffuse models for sampling and predictive inference," *Annals of Statistics, 6*, 1318-1336.

Lane, D., and Sudderth, W. (1984). "Coherent predictive inference," *Sankhyá Ser. A, 46*, 166-185.

Lenk, P. (1984), *Bayesian Nonparametric Predictive Distributions*, Doctoral Dissertation, The University of Michigan.

Lewins, W. A., and Joanes, D. N. (1984), "Bayesian estimation of the number of species," *Biometrics 40*, 323-328.

Lindley, D., and Smith, A. F. M. (1972), "Bayes estimates for the linear model," *Journal of the Royal Statistical Society, Series B*, 34, 1-41.

Luce, R. D., Narens, L. (1987), "Measurement scales on the continuum," *Science, 236*, 1527-1531.

Mandelbrot, B. B. (1982), *The Fractal Geometry of Nature*, W. H. Freeman and Co., New York.

Poincaré, H. (1912), *Calcul des Probabilités*, Deuxiéme Edition, Gauthier-Villars.

Ramakrishnan, S. and Sudderth, W. (1988), "A sequence of coin- toss variables for which the strong law fails." *American Mathematical Monthly, 95*, 939-941.

Rényi, A. (1970), *Probability Theory*, New York : American Elsevier.

Russell, Bertrand (1914), *Our Knowledge of the External World*, Lecture 4, Allen and Unwin: London.

Savage, L. J. (1972), *The Foundations of Statistics*, Second Revised Edition, New York: Dover.

Schervish, M., Seidenfeld, T., and Kadane, J., (1984), "The extent of non-conglomerability," *Z. f. Wahrscheinlichkeitstheorie, 66*, 205-226.

Whitehead, A. N. (1920), *The concept of nature*, Cambridge University Press, Cambridge.

Whitrow, G. J. (1980), *The Natural Philosophy of Time, Second Edition*, Oxford University Press.

Zabell, S. L. (1982), "W. E. Johnson's sufficientness postulate," *The Annals of Statistics, 10*, 1091-1099.